

COMPARISON OF MISSING DATA METHODS USING REGISTER-BASED AUXILIARY DATA FOR HEALTH-RELATED SURVEY DATA OUTCOME

Oona Pentala, Tommi Härkänen and Risto Kaikkonen

National Institute for Health and Welfare, Finland, firstname.lastname@thl.fi

Nonresponse is a ever-progressing issue with population surveys. However, population-based surveys are a vital source of information, because registers contain only limited range of information. By combining survey and register data and using effective missing data methods it is possible to gain more reliable results using survey data suffering from nonresponse issues.

In 2010 Finnish Regional Health and Well-being Study (ATH survey) was conducted by National Institute for Health and Welfare with a national sample of 5,000 and regional samples from Kainuu (9,000), Northern Ostrobothnia (8,000) and city of Turku (9,000). The response rate varied from 37 % to 65 % between different age groups and regions. Overall in the whole data set (n= 31,000) the response rate was 49 %.The study included questions mainly concerning health and well-being. We used this data to compare different missing data handling methods such as Inverse probability Weighting (IPW, used for unit nonresponse), weighted sequential Hot Deck imputation (WSHDI) and multiple imputation (MI) which work with MAR or MNAR data. When using these methods it was also possible to see how they affect the survey results compared to results using only sample weights (CC analysis).

As an outcome in this work we used the self-reported depression which had one of the highest item nonresponse rates in the survey. The question was “Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months?” and depression was one of the alternatives in the question. Because of the question structure and disposition of the alternative, also valuable information was gained from the response indicators of the previous and the following alternative. The overall item nonresponse rate for the self-reported depression was 9.4 % among the participants of the survey. The outcome prevalences of self-reported depression were studied in two age groups 25 to 64-year-olds and over 65 year-olds.

To gain information of the nonrespondents and to develop better nonresponse models we used also register data from National Register Center (sample variables such as age, gender and marital status), Statistics Finland (education and profession of the respondent) and Social Insurance Institute (special reimbursement of medication such as psychotropic medication). The register data was available for the full sample.

Overall the nonresponse bias corrected results of self-reported depression prevalences were quite similar in CC analysis, IPW and WSHDI. MI rates stand out significantly especially in over 65-year-olds, which shows that also the non-complete predictors affect the results. All the methods seemed to respect the relation between areas and age groups compared to CC analysis but also compared to the register-based rates of people receiving reimbursement of depression medication. Overall the results indicate that CC analysis underestimates the rate of depression in every area studied. Correcting this nonresponse bias is vital especially if there is further analysis to be done with depression in the analysis model. The results also suggest that the willingness to answer the depression question depends on depression itself which means that the data is MNAR and the nonresponse corrected estimates may still be biased. Also further examination especially of the MI predictor sets should be done.

References

Kaikkonen R., Murto J., Pentala O., Koskela T., Virtala E., Härkänen T., Koskenniemi T., Ahonen J., Vartiainen E. & Koskinen S. Results of Regional Health and Well-being Study 2010-2014. Internet publication: www.thl.fi/ath. 2010-2014.

Härkänen T., Kaikkonen R., Virtala E. and Koskinen S. Inverse probability weighting and doubly robust methods in correcting the effects of non-response in the reimbursed medication and self-reported turnout estimates in the ATH survey. doi:10.1186/1471-2458-14-1150. BMC Public Health 2014, 14:1150.

Pentala O. Väestötutkimusaineiston tilastolliset kadonhallintamenetelmät - Alueellisenveys- ja hyvinvointitutkimuksen kyselyaineisto 2010 (Statistical Nonresponse Methods in Population Surveys - Regional Health and Wellbeing study 2010). Master's Thesis. University of Helsinki. <http://hdl.handle.net/10138/144287>, 2014.